

TRAVAIL DIRIGÉ DE BIOSTATISTIQUE

www.nursunity.ml

TD Groupe 1

PLAN

- Énoncé
- Question/Solution:

1. Calculer la moyenne et l'écart-type des durées de vie des composants de l'échantillon
2. En déduire un intervalle de confiance de la durée de vie moyenne m des composants avec un coefficient de confiance de 95%
3. Interprétation de résultat
4. Sans changer la taille de l'échantillon, sur quel paramètre peut-on agir pour réduire l'amplitude de l'IC

ÉNONCÉ

- Une usine produit un type de composants électronique. La durée de vie des composants d'un échantillon de 100 composants pris au hasard est consigné dans le tableau suivant :

Durées de fonctionnement en heures	1800	1900	2000	2100
Effectifs	10	40	30	20

1. CALCULER LA MOYENNE ET L'ÉCART-TYPE DES DURÉES DE VIE DES COMPOSANTS DE L'ÉCHANTILLON

- La moyenne m_e :

C'est la somme des valeurs observées divisée par le nombre total des valeurs observées

$$m_e = \sum \frac{n_i \times x_i}{N}$$

m_e = moyenne
x_i = chaque observation
n_i = 1,2,3… n
N= nombre d' observation

- On a:

x_i = 1800 ; 1900 ; 2000 ; 2100

i = 10 ; 40 ; 30 ; 20

N= 100

- Donc:

$$m_e = \frac{(10\times1800) + (40\times1900) + (30\times2000) + (20\times2100)}{100}$$

m_e= 1960 h

L’écart-type:

Il se calcule comme suit :

$$\sigma = \sqrt{\frac{\sum_i n_i (xi - me)^2}{N}}$$

Application numérique

$$\sigma = \sqrt{\frac{10(1800-1960)^2 + 40(1900-1960)^2 + 30(2000-1960)^2 + 20(2100-1960)^2}{100}}$$

σ= 91,65h

❖Remarque :

Pour calculer l’écart-type on a 2 formules, il est préférable de choisir cette formule parce que on travaille sur l’effectif, afin de ne pas obtenir un écart-type très grand.

2. EN DÉDUIRE UN INTERVALLE DE CONFIANCE DE LA DURÉE DE VIE MOYENNE M DES COMPOSANTS AVEC UN COEFFICIENT DE CONFIANCE DE 95%

- Intervalle de confiance :

C'est un intervalle à l'intérieur duquel peut se situer la vraie valeur du paramètre de la population

- L'échantillon est de taille n≥30 et peut être considérer comme non exhaustif; càd que la taille de l'échantillon est négligeable par rapport à la production totale; on peut donc appliquer les résultats du cours sur l'estimation ponctuelle d'une moyenne.
- Pour avoir un coefficient de 95% il faut avoir :

2F (t)-1=0,95 et donc f(t)= 0,975

Niveau de confiance C	Niveau de risque α	Coefficient critique t
90%	10%	1,645
95%	5%	1,960
99%	1%	2,576

et d'après la table :

on a t=1,96

L’intervalle donc sera calculer comme suit :

$$I_{95} = [m_e - t \frac{\sigma}{\sqrt{n-1}} \; ; m_e + t \frac{\sigma}{\sqrt{n-1}}]$$

Cet intervalle est l’intervalle de confiance de la moyenne m de la population avec le coefficient de confiance demandé.

Application numérique :

$$I_{95} = \left[1960 - 1,96 \frac{91,65}{\sqrt{100-1}} \; ; 1960 + 1,96 \frac{91,65}{\sqrt{100-1}}\right]$$

$$I_{95} = \left[1942;1978\right]$$

3. INTERPRÉTATION DE RÉSULTAT

L'intervalle de confiance (IC) à 95% est un intervalle de valeurs qui a 95% de chance de contenir la vraie valeur du paramètre, avec un risque de 5% de faire une erreur, donc la durée de vie moyenne des composants est comprise entre 1942h et 1978h

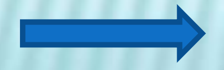
4. SANS CHANGER LA TAILLE DE L'ÉCHANTILLON, SUR QUEL PARAMÈTRE PEUT-ON AGIR POUR RÉDUIRE L'AMPLITUDE DE L'IC ?

- On peut agir sur la taille de l'échantillon on l'augmentant mais selon la question on ne peut pas changer la taille de l'échantillon, donc on va agir sur le coefficient de confiance en le diminuant.
- Supposons que le coefficient de confiance est 90% d'après la table on a t= 1,645

  donc I_{90}= [1945-1975]

  alors l'amplitude de l'intervalle de confiance =

  1975 – 1945 = 30

 $I_{90} < I_{95}$

TD Groupe 2

PLAN

- Définitions des concepts:
- L'énoncé du problème:
- Solution:

DÉFINITION DES CONCEPTS:

a. Intervalle de confiance : on cherche à connaître les valeurs de certaines caractéristiques d'une variable aléatoire grâce à des observations réalisées sur un échantillon.

b. Loi binomiale : Soient les épreuves répétées et indépendantes d'une même expérience de Bernoulli. Chaque expérience n'a que deux résultats possibles : succès ou échec.

c. Loi de Bernoulli: [expérience n'ayant que deux résultats possibles]

par exemple succès et échec. =>la variable aléatoire X qui associe:

la valeur 0 à l'échec (ou à l'absence de la caractéristique)

la valeur 1 au succès (ou à la présence de la caractéristique).

SUITE

d. Échantillon:

C'est une partie de la population qui permet l'étude de la variabilité des caractéristiques d'intérêt de la population, il faut qu'il soit convenablement sélectionné.

e. Loi normale: La distribution normale, ou de Laplace-Gauss, appelée aussi gaussienne, est une distribution continue qui dépend de deux paramètres μ et σ. On la note N(μ, σ^2):

$$f(x;\ \mu,\ \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2}$$

SUITE

f. loi normale centrée réduite:

On dit que la distribution est centrée si son espérance est nulle ; elle est dite réduite si sa variance (et son écart-type) est égale à 1. La distribution normale centrée réduite N(0, 1).

PROBLÈME:

- On veut étudier la proportion p de gens qui vont au cinéma chaque mois.

On prend donc un échantillon de taille n= 100.

Soit N le nombre de personnes dans l'échantillon qui vont au cinéma mensuellement.

Questions

1. Quelle est la loi N ? Par quelle loi peut on l'approcher et pourquoi? En déduire une approximation de la loi de F = N/n ?
2. On observe une proportion f de gens qui vont chaque mois au cinéma. Donner la forme d'un intervalle de confiance p , de niveau de confiance 1- α.
3. Applications numériques : f= 0.1 , 1-α = 90%, 95%,98%.

SOLUTION:

- 1)on suppose que les personnes ont bien été interrogées indépendamment. Ainsi, on a un schéma de Bernoulli : une personne interrogée va au cinéma chaque mois → SUCCES, sinon, ECHEC. Et donc N suit une loi binomiale B(n=100,p)

comme $n \geq 20$, si $np>5$ et $n(1-p)>5$ (à vérifier lors de l'application numérique), et donc F suit approximativement la loi N $\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

SUITE

2)Intervalle de confiance:

$$IC=[f-z(\alpha/2)\sqrt{\frac{P(1-p)}{n}}\ ,\ f+z(\alpha/2)\sqrt{\frac{P(1-p)}{n}}]$$

où $P[Z \geq z(\alpha/2)]=\alpha/2$,Z de loi normale centrée réduite, $1-\alpha$ est le niveau de confiance.

SUITE

3) $f = 0,1$

- 1-α=90% , z(α/2)=1,645 , IC [0.05,0.15]
- 1-α=95% , z(α/2)=1,96 , IC[0.04,0.16]
- 1-α=98% , z(α/2)=2,326 , IC[0.03,0.17]

TD Groupe 3

PLAN

- Introduction
- L’énoncé d’exercice
- La solution
- conclusion

INTRODUCTION

Dans le cadre de notre formation nous avons chargé autant que des étudiants S5 à l'ISPITSE d'approfondir notre connaissance en biostatistique à travers les travaux dirigés, en se concentrant sur un ensemble de sujets dont on a choisi de traiter un exercice intitulé **: l'intervalle de confiance .** cette matière était enseigné par **Mr Hro ougni**

EXERCICE

L'énoncé:

Une entreprise d'import-export gère un parc de 290000 conteneurs. Sur 60 conteneurs pris au hasard, 9 doivent être réparés.

Les questions:

1. Donner une estimation ponctuelle du pourcentage de conteneurs devant être réparés.
2. Déterminer un intervalle de confiance de la proportion de conteneurs qui doivent être réparés avec un risque de 2 %. Donner une interprétation du résultat.
3. Au sein de l’entreprise, on souhaite connaitre la proportion de conteneurs ne nécessitant pas de répartition à ±1 % avec un coefficient de confiance de 99%. Déterminer la taille minimale d’un échantillon permettant d’atteindre cet objectif.

SOLUTION

1. L'estimation ponctuelle se fait à l'aide d'un estimateur, qui est une variable aléatoire d'échantillon. L'estimation est la valeur que prend la variable aléatoire dans l'échantillon observé.

L'estimation ponctuelle du pourcentage de conteneurs devant être réparés est :

$$P_e = 9/60 = 15\%.$$

SOLUTION

2. l'échantillon est de taille n supérieure à 30 et peut être considéré comme non exhaustif (la taille de l'échantillon est négligeable par rapport au parc de conteneurs). On peut donc appliquer le résultat du cours.

   Pour avoir un coefficient de confiance 98%, il faut avoir 2F(t) – 1 = 0.98 et donc F(t) = 0.99. d'après la table, on a t=2.33

L'intervalle $I_{98\%} = \left[Pe - t\sqrt{\frac{Pe(1-Pe)}{n-1}} \; ; \; Pe + t\sqrt{\frac{Pe(1-Pe)}{n-1}} \right]$

Est l'intervalle de confiance de la proportion p des conteneurs doivent être réparés.

Ceci donne :

$I_{98\%} = \left[0.15 - 2.33\sqrt{\frac{0.15(1-0.15)}{60-1}} \; ; \; 0.15 + 2.33\sqrt{\frac{0.15(1-0.15)}{60-1}} \right]$

= [4,17 % ; 25.83 %]

SOLUTION

La phrase suivante est vraie avec une probabilité de 98 % : la proportion de conteneurs qui doivent être réparés est comprise entre 4.17 % et 25.83 %.

SOLUTION

3. Pour connaitre la proportion de conteneurs devant être réparés à ±1 % , il faut un intervalle de confiance d'amplitude 2 %.

Pour avoir un coefficient de confiance de 99 % , il faut avoir 2F(t) – 1 = 0.99 autrement dit F(t) = 0.995. D'après la table, on doit avoir t = 2.58.

pour atteindre les objectifs demandés, il faut avoir un échantillon de taille

$$n \geq 1 + \frac{2.58^2}{0.02^2} = 16\,642$$

conteneurs.

CONCLUSION

En totalité, nous pouvons conclure que ce travail a été bénéfique sur plusieurs niveaux, surtout en ce qui concerne l'application de notre connaissances théorique et affronter sur un plan réel.

TD Groupe 4

PLAN

- Exercice.
- Questions/Réponses.

EXERCICE

- Des citrons sont produits dans des conditions reproductibles par une entreprise agroalimentaire de Sud de L'Espagne pour laquelle vous travaillez. Ces citrons forment une population de référence. Leurs diamètre est distribuer normalement dans cette population avec une moyenne de 7,0 cm et un écart-type de 1,0 cm.
- Un dispositif performant permet également de détecter , sur chaque citrons , la concentration du pesticide absorbé par l'écorce. Cette grandeur est, elle aussi, distribuer normalement dans la population référence avec une moyenne de 2,5 mg/ml et un écart type de 0,2 mg/ml.
- Les citrons sélectionnés pour la vente sont ceux dont le diamètre est compris entre 5,5 et 9,0 cm (inclus) et dont la concentration de pesticide absorbé par l'écorce est inférieur ou égale à 2,8 mg/ml.

QUESTIONS A

- A- calculez la proportion des citrons sélectionnés pour la vente dans la population référence?

Réponse

- A- la proportion de citrons sélectionnés pour la vente dans la population référence (population infinie) correspond à la probabilité $P[(5,5<X<9,0)\cap(Y<2,8)]$;

les deux variables X et Y étant indépendantes:

$P[(5,5<X<9,0)\cap(Y<2,8)] = P(5,5<X<9,0) \times P(Y<2,8)$

En appliquant le changement de variable $W=(X-7,0)/1,0$ et $Z=(Y-2,5)/0,2$:

$P[(5,5<X<9,0)\cap(Y<2,8)] = P(5,5<X<9,0) \times P(Y<2,8)$

$= P(5,5-7,0<W<9,0-7,0) \times P[(Z<(2,8-2,5)/0,2]$

$= P(-1,5<W<2,0) \times P(Z<1,5)$

$= [1-P(W>1,5) + P(W>2,5)] \times [1-P(Z<1,5)]$

$= (1-0,0668 +0,0228) \times (1-0,0668)$

$= 0,9104 \times 0,9332 = 0,85$

QUESTION B

- B- un technicien a mesuré les diamètres et la concentration de pesticide absorbé par l’écorce sur un échantillon de citrons quelle a prélevée de la production journalière.

  Les valeurs sont reportées ci-dessous:

Diamètre (cm)	C6oncentration pesticide (mg/ml)
4,3	2,7
4,6	2,3
4,7	2,6
5,2	2,7
5,4	3,0
5,8	2,8
6,0	2,4
6,1	2,6
6,1	2,6
6,2	2,7
6,2	3,1
6,5	2,1
6,5	2,7
6,6	2,5
6,7	2,2
6,7	2,6
6,8	2,7

Diamètre cm	Concentration pesticide mg/ml
6,8	2,9
6,9	2,6
6,9	2,6
6,9	2,5
7,0	2,7
7,2	2,5
7,3	3,0
7,4	2,6
7,4	2,8
7,7	2,9
7,7	2,7
8,0	2,3
8,3	2,5
8,4	2,5
8,5	2,7
9,0	3,0
9,2	2,6
9,4	2,4

Quelle conclusion vous inspirent ces données ??

RÉPONSES

Population référence: citrons produits dans des conditions reproductibles par la firme agroalimentaire (population infinie).

Soit X la variable aléatoire (quantitative continue): ’’diamètre des citrons en cm’’

X suit à une loi N(7,0;1,0)

Soit Y la variable aléatoire (quantitative continue): ’’ concentration de pesticide absorbée par l’écorce d’un citron en mg/ml’’

Y suit une loi N(2,5; 0,2)

Suite

- B- Il s'agit de réaliser un test de conformité d'une proportion observé a une proportion exacte. Posons Ho : l'échantillon prélevé est issu de la population référence pour laquelle la proportion de citrons sélectionnées pour la vente est ∏= 0,85 (calculer en « A » caractérisant la population référence des citrons produit par la firme).

  La variable d'échantillonnage Po = « proportion de citrons sélectionnées pour la vente dans un échantillon de 35 citrons» subit, sous l'hypothèse nulle Ho, des fluctuation d'échantillonnage de matière binomiale, approchable (N∏=29,75 et N(1-∏) = 5,25 ,sont supérieurs à 5)par la loi normale N(0,85;0,06)

Calcul de Po:

le tableau suivant permet d'établir que 23 citrons sont bons pour la vante dans un échantillon de 35 citrons. Po= 23/35

Citrons pour la vante:

Diamètre (cm)	Concentration pesticide (mg/ml)	Sélection du citron
4,3	2,7	Non
4,6	2,3	Non
4,7	2,6	Non
5,2	2,7	Non
5,4	3,0	Non
5,8	2,8	Oui
6,0	2,4	Oui
6,1	2,6	Oui
6,1	2,6	Oui
6,2	2,7	Oui
6,2	3,1	Non
6,5	2,1	Oui
6,5	2,7	Oui
6,6	2,5	Oui
6,7	2,2	Oui
6,7	2,6	Oui
6,8	2,7	Oui

Diamètre (cm)	Concentration pesticide mg/ml	Sélection du citron
6,8	2,9	Non
6,9	2,6	Oui
6,9	2,6	Oui
6,9	2,5	Oui
7,0	2,7	Oui
7,2	2,5	Oui
7,3	3,0	Non
7,4	2,6	Oui
7,4	2,8	Oui
7,7	2,9	Non
7,7	2,7	Oui
8,0	2,3	Oui
8,3	2,5	Oui
8,4	2,5	Oui
8,5	2,7	Oui
9,0	3,0	Non
9,2	2,6	Non
9,4	2,4	Non

SUITE

Le critère de test est $\varepsilon_o = (P_o - 0{,}85) / 0{,}06$, qui donne $\varepsilon_o=3{,}20$ pour $P_o=0{,}66$ (23/35), soit $\alpha_o=0{,}14\%$ pour un test bilatéral (lecture de la table de la loi normale centrée réduite). Cette valeur de α_o est très inférieure au risque seuil standard $\alpha=5\%$.

Conclusion du test :

On rejette donc Ho avec un risque de 1ère espèce très faible, pratiquement nul. Les différences observées sur la proportion de citrons « bons pour la vente » entre la population référence et l'échantillon sont significatives au risque seuil $\alpha=5\%$ (et même au risque de 1% !). Origine possible des différences :

- mesures mal réalisées par le technicien ;
- quelque chose dans la production a changé, les citrons ne sont plus les mêmes ;
- NP est trop proche de 5 pour une approximation normale confortable ;
- l'échantillon n'a pas été tiré au hasard (non représentatif de la population) ;
- etc....

Question C

- C- Intrigué par ces résultats, vous avez calculé la moyenne et l'écart type des diamètres et des concentrations de pesticide absorbé par l'écorce que vous avez mesuré sur un échantillon de 50 citrons prélevée ou hasard de la production journalière. vous obtenez les valeurs suivantes:

Diamètre:

- moyenne: 6,8 cm
- écart type : 1,1 cm

Concentration de pesticide absorbé par l'écorce:

- moyenne: 2,6 mg/ml
- écart type: 0,2 mg/ml

A quelle conclusion aboutissez-vous finalement avec ce 2éme échantillon?

RÉPONSES

- C- Il s'agit cette fois d'effectuer un test de conformité d'une moyenne observée à une moyenne référence exacte, ce pour les 2 variables X et Y.

Pour ce 2ème échantillon, chaque moyenne observée est accompagnée de son écart type σ_0 . Mais on ne se sert pas de σ_0 dans ces tests sur la moyenne !

En effet : μ et σ sont connus pour les 2 variables dans la population référence (attention, la variable d'échantillonnage n'est pas la variable d'étude X mais sa moyenne qui a pour variance σ^2 / n ; de même concernant la variable Y).

X étant distribué normalement dans la population, sa moyenne l'est également.

SUITE

Comme l'écart type σ exact est connu, on utilise le critère $\text{A}_\circ = |(\bar{X}_\circ - \text{A})|/\text{A}\sqrt{N}$ qui suit une loi normale centrée réduite. Le raisonnement est le même concernant Y et sa moyenne observée.

- Premier test :
- Test sur :

Le critère de test est $\text{A}_\circ = |(6{,}8-7{,}0)|/0{,}5\text{x}\sqrt{50}$

ε_o = 1,41 , qui donne α_o = 15,7 % pour un test bilatéral (lecture de la table de la loi normale centrée réduite). Cette valeur de α_o est très supérieure au risque seuil standard α=5%.

Conclusion du test :

On accepte Ho Les différences observées pour la moyenne entre la population référence et l'échantillon sont imputables au hasard des fluctuations d'échantillonnage au risque seuil α=5% (et même au risque de 10% !). La moyenne observée est conforme à celle de la population référence.

Deuxième test :

Test sur Y_o :

Le critère de test est $\text{A}_\circ = |(2{,}6-2{,}5)|/0{,}2\text{x}\sqrt{50}$

Suite

- εo = 3,54 , qui donne αo = 0,04 % pour un test bilatéral (lecture de la table de la loi normale centrée réduite). Cette valeur de αo est très inférieure au risque seuil standard α=5%, elle est même quasi nulle!

Conclusion du test :

- On rejette Ho Les différences observées pour la moyenne des concentrations de pesticide entre la population référence et l'échantillon sont significatives au risque seuil α =5% (et même au risque de 1% !).

Origine possible des différences constatées sur le 1er échantillon :

- Ce deuxième test confirme celui réalisé en B/ : quelque chose ne va plus dans la production des citrons !

  La moyenne observée pour le diamètre des citrons du 2ème échantillon est pratiquement la même que pour le 1er échantillon,

Suite

- **le doute concerne donc la variable Y :** "concentration de pesticide absorbée par l'écorce d'un citron en mg/ml", qui explique les différences observées sur le premier échantillon et l'échec du test de conformité réalisé en b/.
- Si l'on admet que le technicien a bien réalisé ses mesures sur un échantillon représentatif (Question 1-b/), la concentration de pesticide présente dans l'écorce des citrons est significativement différente de celle constatée dans la population référence (un test de conformité de la moyenne observée de Y pour le premier échantillon, donne $\varepsilon_o = 2{,}96$ soit $\alpha_o = 0{,}31\%$!!!).
- La concentration de pesticide absorbée par l'écorce des citrons est trop forte par rapport à la population référence. Il faudrait d'urgence faire une enquête sur l'épandage du pesticide sur les citrons

TD Groupe 5

Problème

Le pique nique de la fête paroissiale vient de se dérouler comme chaque année.

Cependant ; 55 des 105 personnes ayant participe et que l'on a pu interroger ont présente des symptômes de gastro-entérite la nuit suivante.

L'interrogatoire a porte sur les aliments que les participants avaient mange au cours du pique-nique et sur la survenue éventuelle de symptômes.

Essayez de déterminer a partir du tableau suivant si l'on peut incriminerons des aliments servis au cours du pique-nique commine étant la source de cet épisode

Aliment	Ont consommé		N'ont pas consommé	
	malades	sains	malades	sains
Poulet frit	42	39	13	11
Haricots au four	32	41	23	9
Salade de pommes de terre	51	16	4	34
Pommes chips	36	33	19	17
Thé glacé	47	44	8	6
Café	15	16	40	34
gâteau	32	28	23	22
Jambon cuit	39	37	16	13

Solution du problème

Il s'agit d'un problème d'association (ou d'indépendance)entre deux variables :un aliment; consommé ou pas et statut malade/non malade

La comparaison va donc porter sur deux variables dichotomiques observées sur l'échantillon des sujets ayant participés au pique-nique et que l'on a pu interroger (on va supposer que beaucoup de participants ont été interrogés et que l'on a donc ici une cohorte rétrospective)

L'hypothèse nulle H0 est :il n'y pas de liaison entre la consommation d'un aliment et la survenue de la gastro-entérite.

Le choix du test statistique se porte ici tout naturellement sur le Chi-carré d'indépendance dont on va devoir s'assurer pour chaque comparaison à effectuer que ses conditions d'application sont bien vérifiées (quel que soit l'effectif théorique T , $T \geq 5$)

On choisit un seuil de décision α = 5 % et une formulation bilatérale du test . Pour chaque test effectué , on rejettera H0 dès que le paramètre chi-carré calculé sera supérieur ou égal à 3.84 (ddl =1 car les tableaux de contingence auront tous deux lignes et deux colonnes).

Il convient d'adopter ici une démarche véritablement épidémiologique et donc d'éviter de rendre trop systématique l'utilisation des testes statistiques.

Cette tentation est d'autant plus difficile à éviter que les données ont été informatisées.

En effet , dans ce cas rien n'est plus facile que de faire exécuter par le logiciel la procédure de calcul du test du chi-carré pour tous les aliments sans exception et de « voir ce que donnent les résultats « .pourtant , on ne doit pas effectuer dans cet exemple tous les tests recherchant pour la gastro-entérite une liaison avec chaque aliment proposé au cours du pique-nique.

On peut donner plusieurs raisons pour lesquelles cette stratégies serait incorrecte :

1) plus on effectue de tests statistiques sur échantillon et plus on a de chances de tomber sur une différence statistiquement significative du seul fait du hasard;

2) un test statistique ne donne une information que sur la stabilité, ou la constance, de l'association étudiée et en aucun cas sur son sens (les sujets ayant consommé l'aliment ont-ils été plus souvent malades que ceux qui ne l'ayant pas consommé , ou l'inverse ?);

3) le résultat du test du chi-carré est directement proportionnel aux effectifs sur lesquels il est calculé .

Il convient donc ici d’abord de calculer pour chaque aliment le taux d’attaque chez les consommateurs et les non-consommateurs, d’observer la différence numérique entre ces deux taux, puis de comparer à l’aide du test statistique seulement celles qui sont intéressantes du point de vue épidémiologique , c’est-à-dire celles pour lesquelles la différence de taux est la plus importante et irait dans le sens de la causalité .

Si l’on note au cours de cette phase descriptive de l’analyse une différence importante mais suggérant l’effet protecteur d’un aliment donné , il faut réfléchir à sa signification épidémiologique éventuelle avant de rechercher sa signification statistique.

Le tableau suivant présente pour chaque aliment le taux d’attaque (%) chez les consommateurs et les non-consommateurs.

	Ont consommé			N'ont pas consommé		
Aliment	Malades	Total	TA (%)	Malades	Total	TA (%)
Jambon cuit	39	76	51.3	16	29	55.2
Poulet frit	42	81	51.8	13	24	54.2
Haricots au four	32	73	43.8	23	32	71.9
Pdt	51	67	76.1	4	38	10.5
Pommes chips	36	69	52.2	19	36	52.8
Thé glacé	47	91	51.6	8	14	57.1
Café	15	31	48.4	40	74	54.0
Gâteau	32	60	53.3	23	45	51.1

la stratégie précédemment exposée nous amène à ne retenir a priori qu'un seul aliment pour effectuer le test statistique : la salade de pommes de terre.

Pour les autres aliments , les différences de taux d'attaque sont très faibles et on voit mal alors comment on pourrait incriminer l'un d'entre eux dans la survenue de la gastro-entérite .

Il est par ailleurs intéressant d'observer que c'est pour la salade de pommes de terre que la différence de taux d'attaque est la plus forte (65.6%) mais également que le taux d'attaque est le plus élevé dans le groupe des consommateurs(76.1%) .

Par ailleurs, les consommateurs de salade de pommes de terre représentent prés de deux tiers des personnes interrogatoire (67/105=63.8%).

Ces trois conditions réunies font que, si la différence observée est statistiquement significative , la relation causale sera plus facile à établir.

Il existe une autre différence de taux d'attaque important entre les deux groupes ,pour les haricots au four, mais elle va dans le sens protecteur pour les consommateurs par rapport aux non-consommateurs(-28.1%).

Ceci va être difficile à expliquer surtout été le fait de ceux qui n'ont pas mangé de salade de pommes de terre cette information n'est pas disponible pour l'exercice.

Par conséquence ,nous constituerons en priorité un seul tableau de contingence ,celui destiné à tester l'association éventuelle entre la consommation de la salade de pomme de terres et la survenue de la gastro-entérite.

	Malade	sains	Total
Ont consommé	51	16	67
N’ont pas consommé	4	34	38
Total	55	50	105

Le calcule du paramètre chi-carré est possible car le plus petit effectif théorique ,situé à l’intersection de la deuxième ligne et de la deuxième colonne est suffisamment grand.

$$T2.2 = \frac{38 \times 50}{105}$$

=18.1

r

On utilisera la formule simplifiée du chi-carré de Pearson

$$x^2 = \frac{(a \times b - b \times c)^{2 \times n}}{L1 \times L0 \times C1 \times C0}$$

$$x^2 = \frac{(51 \times 34 - 16 \times 4)^{2 \times 105}}{67 \times 38 \times 55 \times 50}$$

$$x^2 = 41.82$$

Valeur très supérieure au seuil de 3.84 pour un risque de 5% et un degré de liberté.

La liaison statistique est très hautement significative .

On peut rejeter l'hypothèse nulle avec un risque p qui ne dépassera pas 0,001

conclure qu'il existe une liaison entre la consommation de salade de pommes de terre et la survenue de la gastro-entérite:

le risque de gastro-entérite est 7,2 fois plus élevé chez les consommateurs (10,5%)et cette différence est statistiquement significative au risque de1pour 1000.

L'enquête épidémiologique peut donc continuer en s'orientant sur ce véhicule de l'épidémie.

Pour information ,la seule autre différence statistiquement significative serait trouvée pour les haricots au four ($X^2 = 7{,}01 : p < 0.01$).

mais comme on l'a dit précédemment, la plus grande prudence est nécessaire dans l'interprétation de ce résultat car il semble difficile d'évoquer un rôle protecteur pour cet aliment

TD Groupe 6

PLAN

- Rappel
- Enoncé
- Solution

A- Loi du χ^2 (chi-2)

C’est une loi dérivée de la loi normale, très importante pour ses applications en statistiques comme nous le reverrons dans les tests.

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes, chacune étant distribuée selon une loi normale centrée réduite : $\forall i, X_i \sim N(0, 1)$
La distribution de $S = X^2 + X^2 + \ldots + X^2$ (somme des carrés des X) est appelée loi de χ^2 à n de grés de liberté (en abrégé d. d. l = degrés de liberté), que l’on note $\chi^2(n)$ où n est le nombre de d. d. l., seul paramètre de la loi.

B- Puissance d’un test

C’est une démarche qui consiste à prendre en compte deux hypothèse synthétiques (hypothèse Nulle et hypothèse Alternative) et tester la probabilité de rejeter HN face à HA en reformulant un problème médical en termes statistiques

EXERCICE

Les premiers éléments d’une enquête sur une épidémie de 83 cas d’hépatite A, et en particulier l’analyse de la courbe épidémique, font envisager l’existence d’une source commune. On observe que 50 des sujets atteint allaient en classe dans le même lycée. Les 50 élevés malades ont été apparies pour l’Age et le sexe avec 50 élèves en bonne santé. On les a tous interrogés sur les possibilités d’exposition à différentes sources d’infection au cours de la période suspecte. On a obtenu les renseignements suivants concernant trois des sources possibles.

source	Cas et Témoins exposés	Cas et Témoins NON exposés	Cas exposé Témoins non exposé	Cas non exposés Témoins exposés
A	14	10	12	14
B	20	3	25	2
C	18	6	9	17

- Déterminer si une des sources peut être incriminée dans la survenue de cette épidémie.

SOLUTION

- Cet exercice s’agit d’un problème d’association entre deux variables : *une source*, présente ou pas, et *l’hépatite A*, présente ou absente, mais cette fois les séries sont appariées .Le test de référence est le test de Chi-carré de Mac Nemar, sous réserve de la vérification de ses conditions d’application (la somme des paires discordantes doit être supérieure ou égale à 10).On se limitera ici à la présentation des deux dernières étapes du test.

1) Si les nombres de paires discordantes « cas exposé - témoin non exposé » (noté en général f) et de paires discordantes « cas non exposé – témoin exposé » (noté en général g) sont très voisins

il y a peu de chance de conclure à une différence statistiquement significative

2) S'il y a plus de paires g que de paires f , l'association que l'on va tester va aller dans le sens inverse de ce que l'on recherche

En présence de la source, le risque de maladie est diminué (il faut alors réfléchir à priori à l'interprétation d'une telle observation, surtout si le résultat du test devait être statistiquement significatif).

1) choix des variables:

-Variable qualitative binaire(source d'infection: présente/absente)

- Variable qualitative binaire(hépatite A : présente/absente)

2) choix de test de référence:

c'est le test de Chi-2 Alors pour tester le rôle possible d'une source, il faut réaliser un tableau de contingence.

3)Tableau de contingence:

A partir des données de problème il est possible de reconstituer les 3 tableaux à 2 lignes et 2coloones comme suit:

Source A

		Témoins		
		Exposé	Non exposé	Total
CAS	exposé	14	12	26 (52%)
	Non exposé	14	10	24 (48%)
Total		28	22	50

Source B

		Témoins		
		Exposé	Non exposé	Total
CAS	exposé	20 (19,8 %)	25 (25,2 %)	45 (90%)
	Non exposé	02 (2,2%)	03 (2,8%)	05 (10%)
Total		22	28	50

Source C

		Témoins		
		Exposé	Non exposé	Total
CAS	exposé	18	09	27 (54%)
	Non exposé	17	06	23 (46%)
Total		35	15	50

- 4) Formulation des hypothèses:

H : une source est l’origine de l’épidémie

HA: l’inverse de HN

On suppose que HN est vrai

On teste la liaison entre les deux variables par toutes les possibilités offertes en l’occurrence des 3 sources concernant l’hypothèse nulle.

Si: f=(cas exposé - témoin non exposé)

Et: g=(cas non exposé – témoin exposé)

- 5) Calcule de Chi-2

On a X² = $\sum_{i,j} \frac{(O-T)^2}{T}$

Donc :

X²(B)= 0,035

X²(C)=2,46

Et d'après le table de lois de Chi-2:

H_o

On a X²(B)<VC (valeur critique) donc P(B) est faible c.-a-d. on peut admettre H H_o

Et

X²(C)>VC donc P(C) est forte c.-à-d. va être rejeter

En définitive. on peut conclure que *la source B* est très probablement a l'origine de l'épidémie. Comme ‘il y a 45 cas sur 50 (90%) qui ont été exposés à la source B, on peut penser que la source B n‘est pas la seule origine de l'épidémie, ou bien qu'une transmission secondaire, par exemple de personne à personne s'est produite.

TD Groupe 7

PROBLÈME N°4:

On fait l'hypothèse que la fertilité des femmes hospitalisée dans le service de médecine générale d'un hôpital diffère de la fertilité de la population générale.

Pour 200 femmes mariées hospitalisées , la distribution de fréquence du nombre d'enfants est présentée dans le tableau ci-dessous, avec pour référence la distribution de fréquence du nombre d'enfants pour les femmes mariées de la population générales correspondante.

Essayez de déterminer si l'hypothèse faite est vérifiée.

➢ DISTRIBUTION DES FRÉQUENCES DU NOMBRE D'ENFANTS

Nombre d'enfants	Nombre de femmes mariées hospitalisées	% de femmes mariées de la population générale
0	56	21.6
1	58	30.3
2	43	26.5
3	20	11.8
4	12	5.0
5 et plus	11	4.8
Total	200	100

SOLUTION DU PROBLÈMES N°4

- La question posée est la suivante:
  - l'observation faite sur l'échantillon est-elle conforme à ce que l'on sait de la fertilité de la population féminine de référence?
- La comparaison porte sur une variable qualitative quelconque, la fréquence du nombre d'enfants, à K modalités (ici, K=6).
- L'objectif est de comparer la distribution observée à la distribution théorique qui a également K modalités.
- L'hypothèse nulle H0 est :
  - les femmes mariées hospitalisées ont la même fertilité que la population générale.

Suite:

- Le test statistique de référence est le Chi-carré de conformité, sous réserve de la vérification de ses conditions d'application (quel que soit l'effectif théorique T,T≥5).

- On choisit un seuil de décision α=5% et une formulation bilatérale du test. Puisque K=6, ddl=5 et la valeur-seuil du Chi-carré est 11.07.

α ddl	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,0158	0,455	1,074	1,642	2,706	3,841	5,412	6,635	10,827
2	0,211	1,386	2,408	3,219	4,605	5,991	7,824	9,210	13,815
3	0,584	2,366	3,665	4,642	6,251	7,815	9,837	11,345	16,266
4	1,064	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,467
5	1,610	4,351	6,064	7,289	9,236	11,070	13,388	15,086	20,515
6	2,204	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457
7	2,833	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322
8	3,490	7,344	9,524	11,030	13,362	15,507	18,168	20,090	26,125
9	4,168	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	4,865	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588
11	5,578	10,341	12,899	14,631	17,275	19,675	22,618	24,725	31,264
12	6,304	11,340	14,011	15,812	18,549	21,026	24,054	26,217	32,909
13	7,042	12,340	15,119	16,985	19,812	22,362	25,472	27,688	34,528
14	7,790	13,339	16,222	18,151	21,064	23,685	26,873	29,141	36,123
15	8,547	14,339	17,322	19,311	22,307	24,996	28,259	30,578	37,697
16	9,312	15,338	18,418	20,465	23,542	26,296	29,633	32,000	39,252
17	10,085	16,338	19,511	21,615	24,769	27,587	30,995	33,409	40,790
18	10,865	17,338	20,601	22,760	25,989	28,869	32,346	34,805	42,312
19	11,651	18,338	21,689	23,900	27,204	30,144	33,687	36,191	43,820
20	12,443	19,337	22,775	25,038	28,412	31,410	35,020	37,566	45,315
21	13,240	20,337	23,858	26,171	29,615	32,671	36,343	38,932	46,797
22	14,041	21,337	24,939	27,301	30,813	33,924	37,659	40,289	48,268
23	14,848	22,337	26,018	28,429	32,007	35,172	38,968	41,638	49,728
24	15,659	23,337	27,096	29,553	33,196	36,415	40,270	42,980	51,179
25	16,473	24,337	28,172	30,675	34,382	37,652	41,566	44,314	52,620
26	17,292	25,336	29,246	31,795	35,563	38,885	42,856	45,642	54,052
27	18,114	26,336	30,319	32,912	36,741	40,113	44,140	46,963	55,476
28	18,939	27,336	31,391	34,027	37,916	41,337	45,419	48,278	56,893
29	19,768	28,336	32,461	35,139	39,087	42,557	46,693	49,588	58,302
30	20,599	29,336	33,530	36,250	40,256	43,773	47,962	50,892	59,703

SUITE

- Pour le calcul du paramètre, il faut tout d'abord calculer l'effectif théorique T de chaque classe pour l'échantillon de 200 femmes mariées hospitalisées en supposant qu'elles sont représentatives de la population générale, c'est-à-dire en appliquant les pourcentages attendus ou théoriques.

- On vérifie bien que tous les effectifs théoriques sont grands. Le fait que ces effectifs ne soient pas des nombres entiers n'est pas gênant car il s'agit bien de valeurs théorique; il importe cependant de ne pas les arrondir pour conserver la précision des calculs suivants.

SUITE

Nombre d'enfants	Fréquence observée	% théorique	Fréquence calculée
0	56	21.6	43.2
1	58	30.3	60.6
2	43	26.5	53.0
3	20	11.8	23.6
4	12	5.0	10.0
5 et plus	11	4.8	9.6
Total	200	100	200

SUITE

- Le calcul du paramètre donne : $\chi^2 = \sum \frac{(O-T)2}{T} = 6.94$
- Le Chi-carré calculé est très inférieur à la valeur-seuil lue dans la table(6,94<11,07).
- On n’est pas en mesure de rejeter H0 et on est amené à conclure que l’échantillon de femmes mariées hospitalisées est bien représentatif de la population générale en ce qui concerne le nombre d’enfants.

SUITE

- Une erreur à ne pas commettre aurait été de vouloir comparer directement des pourcentages en transformant les fréquences observées en pourcentages observés et en les comparant aux pourcentages théoriques à l'aide d'un test de comparaison de pourcentages pris deux à deux.

- En procédant ainsi, on ramènerait les calculs à un échantillon de taille 100 au lieu des 200 effectivement observés et la puissance de cette comparaison en serait d'autant moins bonne.

TD Groupe 8

- Une usine produit un type de composants électroniques. La durée de vie des composants d'un échantillon de 100 composants pris au hasard est consigné dans le tableau suivant :

Durée de fonctionnement (en heure)	1800	1900	2000	2100
Effectifs	10	40	30	20

1. Calculer la moyenne et l’écart type σe des durée de vie des composants de l’échantillon
2. En déduire un intervalle de confiance de la durée de vie moyenne des composants avec un coefficient de confiance de 95 %
3. Donner une interprétation du résultat
4. Sans changer la taille de l’échantillon, sur quel paramètre peut-on agir pour réduire l’amplitude de l’intervalle de confiance.

Solution

1. [illegible] durée de vie des [illegible]

On a :

$$me = \sum_{i=1}^{n} \frac{ni\ xi}{N}$$

$$\text{donc } me = \frac{10\times1800+40\times1900+30\times2000+20\times2100}{100} = 1960h$$

SUITE

L’écart type :

$$\sigma = \sqrt{\frac{\sum_k n_k (x_k - m)^2}{n}}$$

$$\sqrt{\frac{10\times(1800-1960)^2+40\times(1900-1960)^2+30\times(2000-1960)^2+20\times(2100-1960)^2}{100}}$$

Donc σe = 91,65h

SUITE

2 En déduire un intervalle de confiance de la durée de vie moyenne des composants avec un coefficient de confiance de 95 % :

L’échantillon est de taille n supérieure à 30 et peut être considéré comme non exhaustif (la taille de l’échantillon est négligeable par rapport à la production totale) ; on peut donc appliquer les résultats du cours sur l’estimation ponctuelle d’une moyenne .

SUITE

- Pour avoir un coefficient de confiance de 95%, il faut avoir 2F(t)-1=0,95 et donc F(t)=0,975. D'après la table, on a t= 1,96
- L'intervalle I95 = $\left[me - t \frac{\sigma e}{\sqrt{n-1}} \; ; \; me + t \frac{\sigma e}{\sqrt{n-1}} \right]$

est l'intervalle de confiance de la moyenne m de la population avec le coefficient de confiance demandé.

SUITE

On a donc :

$$I95 = \left[1960 - 1{,}96 \times \frac{91{,}65}{\sqrt{100-1}} \; ; \; 1960 + 1{,}96 \times \frac{91{,}65}{\sqrt{100-1}} \right]$$

I95= [1942;1978]

INTERPRÉTATION

3 . avec risque de 5% de faire une erreur, la durée de vie moyenne des composants est comprise entre 1942h et 1978h

SUITE

4. Sans changer la taille de l'échantillon, sur quel paramètre peut-on agir pour réduire l'amplitude de l'intervalle de confiance.

On peut agir sur la taille de l'échantillon en l'augmentant et/ou sur le coefficient de confiance en le diminuant

TD Groupe 9

Problèmatique:

Dans une agence de location de voitures, le patron veut savoir quelles sont les voitures qui n'ont roulé qu'en ville pour les revendre immédiatement.

Pour cela, il y a dans chaque voiture une boite noire qui enregistre les nombre d'heurs pendants lesquelles la voiture est restée au point mort, au premier rapport, au deuxième rapport,…, au cinquième rapport.

On sais qu'une voiture qui ne roule qu'en ville passe en moyen 10% de son temps au point mort, 5% en première, 30% en second, 30% en troisième, 20% en quatrième et 5% en cinquième.

On décide de faire un test du X^2 pour savoir si une voiture n'a roulé qu'en ville ou non.

Question:

1) Sur une première voiture, on constate sur 2000 heures de conduite :

210 h au point mort, 94 h en première, 564 h en seconde, 630 h en troisième, 390 h en quatrième, et 112 h en cinquième. Cette voiture n'a-t-elle fait que rester en ville ?

On veut tester l'adéquation de notre échantillon a la loi discrète :

- $p_0=0.1$
- $p_1=0.05$
- $p_2=0.3$
- $p_3=0.3$
- $p_4=0.2$
- $p_5=0.05$

On effectue un teste du χ^2. En fait, en veut tester **H_0= la voiture n'a roulé qu'en ville, contre H_1= la voiture n'a pas roulé qu'en ville**.

1) Pour la première voiture, on constate:

	0	1	2	3	4	5
eff obs obs_i	210	94	564	630	390	112
eff th th_i	200	100	600	600	400	100

On calcule la distance du X² :

$$D^2 = \sum_{i=0}^{5} \frac{(th_i - obs_i)^2}{th_i} = \frac{10^2}{200} + \frac{6^2}{100} + \frac{36^2}{600} + \frac{10^2}{600} + \frac{10^2}{400} + \frac{12^2}{100} = 6.21$$

Détermination du seuil :

$$P[\chi_5^2 > c] = 0.05 \Longrightarrow c = 11.07.$$

Comme $D^2 = 6.21 < 11.07$, on ne peut rejeter H_0: la voiture n'a roulé qu'en ville.

TABLE DE χ^2

α / ddl	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,0158	0,455	1,074	1,642	2,706	3,841	5,412	6,635	10,827
2	0,211	1,386	2,408	3,219	4,605	5,991	7,824	9,210	13,815
3	0,584	2,366	3,665	4,642	6,251	7,815	9,837	11,345	16,266
4	1,064	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,467
5	1,610	4,351	6,064	7,289	9,236	11,070	13,388	15,086	20,515
6	2,204	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457
7	2,833	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322
8	3,490	7,344	9,524	11,030	13,362	15,507	18,168	20,090	26,125
9	4,168	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	4,865	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588
11	5,578	10,341	12,899	14,631	17,275	19,675	22,618	24,725	31,264
12	6,304	11,340	14,011	15,812	18,549	21,026	24,054	26,217	32,909
13	7,042	12,340	15,119	16,985	19,812	22,362	25,472	27,688	34,528
14	7,790	13,339	16,222	18,151	21,064	23,685	26,873	29,141	36,123
15	8,547	14,339	17,322	19,311	22,307	24,996	28,259	30,578	37,697
16	9,312	15,338	18,418	20,465	23,542	26,296	29,633	32,000	39,252
17	10,085	16,338	19,511	21,615	24,769	27,587	30,995	33,409	40,790
18	10,865	17,338	20,601	22,760	25,989	28,869	32,346	34,805	42,312
19	11,651	18,338	21,689	23,900	27,204	30,144	33,687	36,191	43,820
20	12,443	19,337	22,775	25,038	28,412	31,410	35,020	37,566	45,315
21	13,240	20,337	23,858	26,171	29,615	32,671	36,343	38,932	46,797
22	14,041	21,337	24,939	27,301	30,813	33,924	37,659	40,289	48,268
23	14,848	22,337	26,018	28,429	32,007	35,172	38,968	41,638	49,728
24	15,659	23,337	27,096	29,553	33,196	36,415	40,270	42,980	51,179
25	16,473	24,337	28,172	30,675	34,382	37,652	41,566	44,314	52,620
26	17,292	25,336	29,246	31,795	35,563	38,885	42,856	45,642	54,052
27	18,114	26,336	30,319	32,912	36,741	40,113	44,140	46,963	55,476
28	18,939	27,336	31,391	34,027	37,916	41,337	45,419	48,278	56,893
29	19,768	28,336	32,461	35,139	39,087	42,557	46,693	49,588	58,302
30	20,599	29,336	33,530	36,250	40,256	43,773	47,962	50,892	59,703

2) Avec une autre voiture, on obtient les données suivantes :

- ✓ 220 h au point mort
- ✓ 80 h en première
- ✓ 340 h en seconde
- ✓ 600 h en troisième
- ✓ 480 h en quatrième
- ✓ 280 h en cinquième

- Pour la seconde voiture , on constate

	0	1	2	3	4	5
eff obs obs_i	220	80	340	600	480	280
eff th th_i	200	100	600	600	400	100

On calcul la distance du X².

$$D^2 = \sum_{i=0}^{5} \frac{(th_i - obs_i)^2}{th_i} = 458.67 >> 11.07$$

On rejette H0 :
la voiture n'a pas roulé qu'en ville. La p-valeur vaut 0. la décision ne fait pas de doute.

TD Groupe 10

ENONCÉ DE PROBLÈME

- On a inoculé 15 rats de laboratoire avec un certain germe pathogène.
- Puis, après tirage au sort, huit d'entre eux ont été traités avec un nouveau médicament et sept ont survécu.
- On observe que deux des sept rats du groupe non traité ont survécu.
- Essayez de déterminer si le médicament a réellement une action sur la survie.

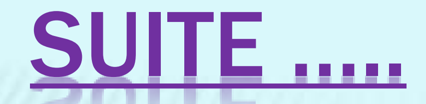
SOLUTION DU PROBLÈME

Pour résoudre ces exercices, il convient de suivre étape par étape la démarche nécessaire à la réalisation d'un test statistique.

On se rappellera que les trois premières étapes, et en particulier la troisième (choix du paramètre qui sous l'hypothèse nulle obéit à une loi de probabilité connue)

SUITE….

- La question posée est : y a-t-il association entre traitement et survie ?
- La comparaison porte sur une variable qualitative dichotomique (survie ou décès).
- L'objectif est de comparer une proportion p1 de souris ayant survécu après traitement (p1 = 7/8 = 87,5 % ; n1 = 8) à une proportion p2 de souris ayant survécu en l'absence de traitement (p2 = 2/7 = 28,6 % ; n2 = 7).
- **L'hypothèse nulle H0 est : il n'y a pas d'association entre traitement et survie, c'est-à-dire p1 = p2.**



- Le choix du test statistique dépend du type de séries à comparer - ici, deux séries indépendantes et de la vérification des conditions d'application.

- Les quantités (n1 * p), (n1 * q), (n2 * p) et (n2 * q) sont toutes inférieures à 5 avec la proportion p = (7+2)/(8+7) = 0,60 et q = 0,40. On est donc amené à retenir le test exact de Fisher comme méthode de comparaison à partir du tableau suivant :

	Survivants	Morts	Total
Traité	7	1	8
Non traité	2	5	7
Total	9	6	15



- On choisit un seuil de décision **α = 5 %**
- On rejettera l'hypothèse nulle si la probabilité calculée par le test exact de Fisher est inférieure ou égale à 0,05.
- On choisira une formulation bilatérale du test car il n'est pas exclu à priori que le traitement ait un effet imprévisible sur la survie, en l'occurrence un effet néfaste.
- L'hypothèse alternative H1 s'écrit donc de façon générale : le traitement a un effet sur la survie.
- Le calcul de la probabilité observée p (a = 7) à partir du tableau précédent s'effectue comme suit :

p1 = 8 ! 7 ! 9 ! 6 ! /7! 1 ! 2 ! 5 ! 15 != 0,0336



- On peut concevoir une configuration encore plus extrême que celle qui a été observée.
- Le tableau suivant représente ce cas de figure où 100% des souris traitées ont survécu, les effectifs des autres cases se déduisant des totaux marginaux restés fixes :

	Survivants	Morts	Total
Traité	8	0	8
Non traité	1	6	7
Total	9	6	15



- Le calcul de la probabilité associée à cette deuxième configuration :

p2 = 8 ! 7 ! 9 ! 6 ! /8 ! 0 ! 1 ! 6 ! 15 != 0,0014

- La configuration la plus extrême dans le sens opposé serait celle où deux seulement des huit souris du groupe traité auraient survécu (2/8 = 25 % de succès).

- Alors, en maintenant fixes les totaux marginaux, le pourcentage de souris non traitées et ayant survécu serait de 100%, et la différence entre les deux groupes, 75 %, serait bien plus importante que celle qui a été observée : 87,5 % - 28,6 % = 58,9 %. Le tableau suivant résume ce troisième cas de figure.

	Survivants	Morts	Total
Traité	2	6	8
Non traité	7	0	7
Total	9	6	15

SUITE….

- La probabilité associée à cette troisième configuration :

  p3 = 8 ! 7 ! 9 ! 6 !/2 ! 6 ! 7 ! 0 ! 15 ! = 0,0056

- Il ne pourrait y avoir de configuration plus défavorable au traitement que cette dernière car l'on n'a observé au total que 6 décès au cours de l'expérience.

- La configuration suivante (3 survies et 5 décès dans le groupe traité) serait en fait moins extrême que la configuration observée car alors la différence de pourcentage de survie entre les deux groupes ne serait plus que de 48,2 %.



- Par conséquent, la probabilité exacte de survenue de la configuration observée ou d'une configuration encore plus défavorable à l'hypothèse nulle du seul fait du hasard est :

$$p = p1 + p2 + p3 = 0{,}0336 + 0{,}0014 + 0{,}0056 = 0{,}0406$$

- Cette probabilité est inférieure à 0,05. On rejette donc l'hypothèse nulle et on est amené à accepter l'hypothèse d'une association entre le traitement et la survie.

- L'observation du sens de la différence de survie entre les deux groupes conduit à conclure à l'efficacité du traitement.

- Si on était parti d'une formulation unilatérale du test, c'est-à-dire en faisant l'hypothèse alternative que le traitement ne pouvait qu'améliorer la survie, la probabilité serait:

$$p = p1 + p2 = 0{,}0336 + 0{,}0014 = 0{,}035$$

CONCLUSION

On remarquera que la simple multiplication par 2 du résultat obtenu avec la formulation unilatérale ne donnerait en rien le résultat de la formulation bilatérale du test.

La formulation bilatérale est finalement plus conservatoire, même si dans le cas présent les deux formulations amènent à rejeter l'hypothèse nulle.

TD Groupe 11

DONNÉES

On a consigne les primes de fin d'année attribuées aux salariés d'une entreprise dans le tableau suivant :

Primes (centaines d'euros)	[0;6[	[6;10[	[10;14[	[14;16[
effectifs	41	79	78	2
Milieux des classes	3	8	12	15

question ° 1

Quelle est la population étudiée ?

Réponse ° 1

la population étudiée est l'ensemble des salariés de l'entreprise considérée

c'est une population qui est composée des salariés des (individus) et ces derniers se représentent au grand nombre

Question °2

Quel est le caractère ?

Réponse°2

ce caractère pouvant prendre toutes les

val… d’… …tère est

le prime est le caractère , est un caractère
quantitatif présente le montant de chaque
individu de cette population et il se diffère
d’une valeur a l’autre donc il est continu

Question °3

Quelle est la nature de ce caractère ?

Réponse °3

Ce caractère pouvant prendre toutes les valeurs d’inter… …est qua…

Le caractère étudié est un variable aléatoire qui ne peut prendre que des valeurs numériques il est dit quantitatif continue car , toutes les valeurs sont possibles, au moins sur un intervalle ex : 41 des salariés ont un prime entre 0 et 6

Question° 4

- Pourquoi a-t-on regroupé les primes en classe ?

Réponse° 4

- Le caractère étant continu , les modalités (valeurs prises par le caractère) sont re s)

Autant que les valeurs des primes sont multiples et la population est nombreuse et pour bien organiser est ordonner la représentation des données, on regroupe les valeurs des primes en des intervalles ou bien des classes

Question°5

Déterminer la moyenne ?

Réponse°5

Notons $\bar{x}$ la moyenne cherchée .pour calculer cette moyenne , on complète le tableau en calculant les milieux des classes

- On a : $\bar{X} = \dfrac{41 \times 3 + 79 \times 8 + 78 \times 12 + 2 \times 15}{41 + 79 + 78 + 2}$

$\bar{\bar{X}} =$ 8,6 centaines d'euros

$$\bar{X} = \frac{\Sigma(\text{effectif x milieu de classe})}{\Sigma \text{ effectif}}$$

Question°6

déterminer l'écart type

Réponse°6

Notons V(x) la variance de la série statistique et б(x) son écart-type . On a:

$$V(x) = \frac{41X(3-8,6)^2+79X(8-8,6)^2+78x(12-8,6)^2+2X(15-8,6)^2}{41+79+78+2}$$

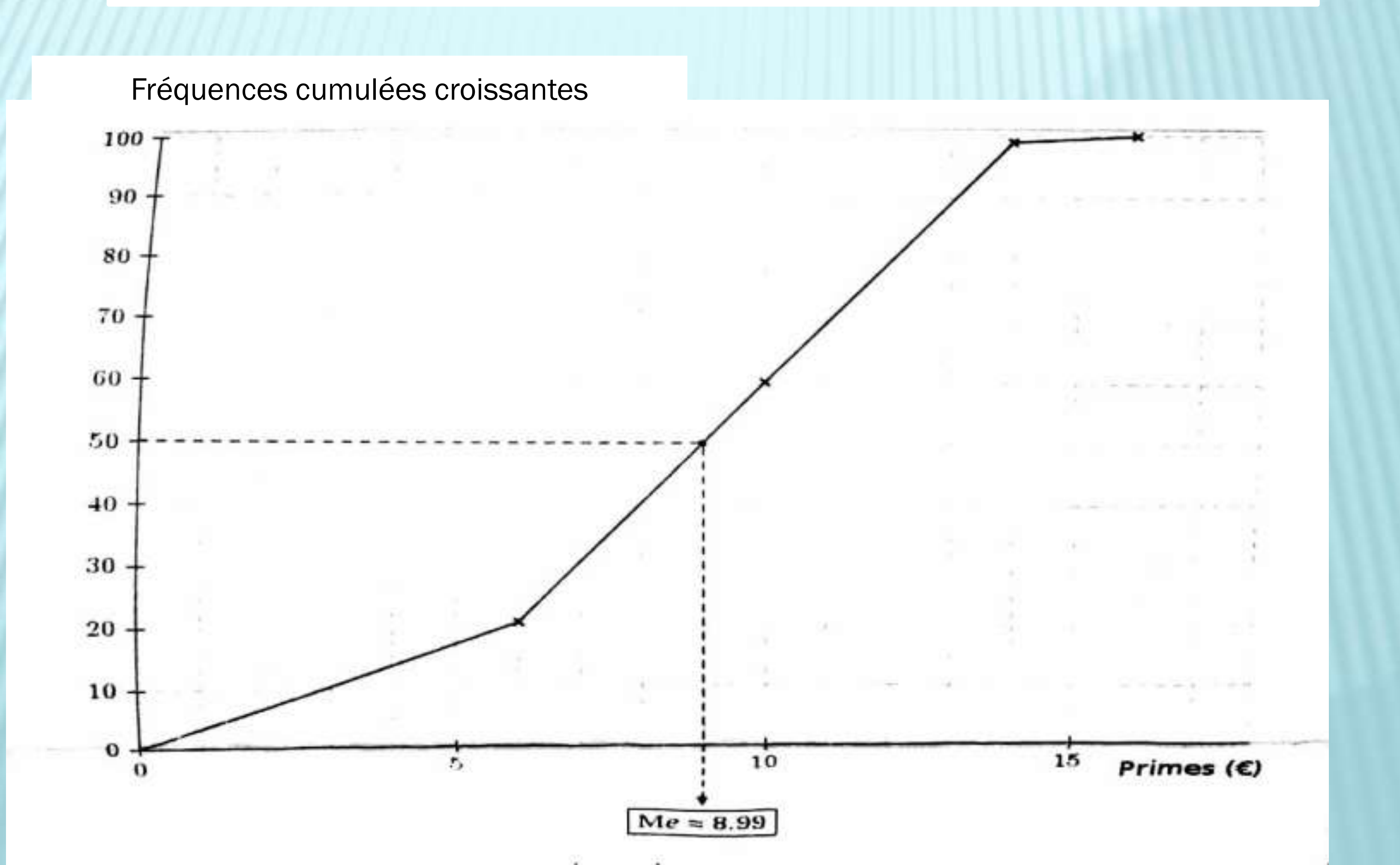
=11,48

On a donc б(x) =√V(x) = √11,48 = 3,38 centaines d'euros

Question° 7

Tracer la courbe cumulative des effectifs

Réponse° 7



Question°8

Déterminer graphiquement la médiane et interpréter .

Réponse°8

On obtient Me≈8,89 centaine d'euros.

- 50% des salariés touchent une prime inférieure à 8,89 centaines d'euros.
- 50% des salariés touchent une prime supérieure à 8,89 centaines d'euros.

TD Groupe 12

EXERCICE:

Dans un centre avicole, des études antérieures ont montré que la masse d'un œuf choisi au hasard peut être considérée comme la réalisation d'une variable aléatoire normale X, de moyenne m et de variance σ. On admet que les masses des œufs sont indépendantes les unes des autres. On prend un échantillon de n=36 œufs que l'on pèse.

Les mesures sont données (par ordre croissant) dans le tableau suivant:

50.34	52.62	53.79	54.99	55.82	57.67
51.41	53.13	53.89	55.04	55.91	57.99
51.51	53.28	54.63	54.12	55.95	58.10
52.07	53.30	54.76	54.24	57.05	59.30
52.22	53.32	54.78	54.28	57.18	60.58
52.38	53.39	54.93	54.56	57.31	63.15

SUITE

a) Calculer la moyenne empirique et l'écart-type empirique de cette série statistique . Tracer le boxplot et un histogramme.

b) Donner une estimation des paramètres m et σ

c) Donner un intervalle de confiance au niveau 95% ,puis 98% ,de la masse moyenne m d'un œuf

SOLUTION:

a) $\bar{x} = 1/n \sum xi = 1982.99/36 = 55.083$;

s=2.683 ;Q1=53.29 ; Med= 54.96 Q3=56.5.

- **Boxplot:** moust1 =50.34 ; moust 2= 60.58

un outlier =63.15

- **histogramme**

	Effectif	Largeur	hauteur
50-52	3	2	1.5
52-54	11	2	5.5
54-56	13	2	6.5
56-58	5	2	2.5
58-64	4	6	0.67

suite

b) $\bar{x}$ est une estimation de m , s est une estimation de σ.

c) IC de niveau de confiance 1-α=95% pour m:

$[\bar{x} - z_{\alpha/2}\, s/\sqrt{36},\ \bar{x} + z_{\alpha/2}\, s/\sqrt{36}] = [54.207, 55.96]$

car $z_{\alpha/2} = z_{0.025}$, $P[Z \leq 1.96] = 0.975$ quand Z de loi N(0,1), et donc $z_{\alpha/2} = 1.96$

IC de niveau de confiance 1-α=98% pour m

$[\bar{x} - z_{\alpha/2}\, s/\sqrt{36},\ \bar{x} + z_{\alpha/2}\, s/\sqrt{36}] = [54.043, 56.123]$

car $z_{\alpha/2} = z_{0.001}$, $P[Z \leq 2.3263] = 0.99$ quand Z de loi N(0,1), et donc $z_{\alpha/2} = 2.3263$

TD Groupe 13

ENONCÉ DE TD

Une entreprise d'import-export gère un parc de 290000 conteneurs. Sur 60 conteneurs pris au hasard ,9 doivent être réparés.

1. donner une estimation ponctuelle du pourcentage de conteneurs devant être réparés

2. déterminer un intervalle de confiance de la proportion de conteneurs qui doivent être réparer avec un risque de 2 % . donner une interprétation du résultat.

3. au sein de l'entreprise , on souhaite connaître la proportion de conteneurs ne nécessitant pas de réparation à ± 1% avec une coefficient de confiance de 99% Déterminer la taille minimale d'un échantillon permettant d'atteindre cet objectif .

CORRECTION DE TD

1. $p_e = \frac{9}{60} = 15\%$

2. L'échantillon est de taille n supérieure à 30 et peut être considère comme non exhaustif (la taille de l'échantillon est négligeable par rapport au parc de conteneurs) . On peut donc appliquer les résultats du cours.

Pour avoir un coefficient de confiance de 98 % il faut avoir $2F(t) - 1 = 0.98$ et donc $F(t) = 0,99$.

d'après la table, on a $t = 2,33$.

L’intervalle $I_{98\%} = \left[pe - t\sqrt{\frac{pe(1-pe)}{n-1}} \,;\, pe + t\sqrt{\frac{pe(1-pe)}{n-1}} \right]$

Est l’intervalle de confiance de la proportion p des conteneurs devant être réparé .

Ceci donne:

$$\mathbf{I_{98\%}} = \left[0,15 - 2,33\sqrt{\frac{0,15(1-0,15)}{60-1}} \,;\, 0,15 + 2,33\sqrt{\frac{0.15(1-0,15)}{60-1}} \right]$$

$$[4,17\% \,;\, 25,83\%]$$

La phrase suivante est vraie avec une probabilité de 98%: la proportion de conteneurs devant être réparés est comprise entre 4,17% et 25,83%.

3. Pour connaître la proportion de conteneurs devant être réparés à ± 1% ,il faut un intervalle de confiance d'amplitude 2%.

Pour avoir un coefficient de confiance de 99%,il faut avoir 2F(t)_1=0.99 autrement-dit F(t)= 0.995. D'après la table , on doit avoir t= 2.58.

Pour atteindre les objectifs demandés, il faut avoir un échantillons de taille:

$$n \geq 1 + \frac{2,58^2}{0,02^2} = 16\,642 \text{ conteneurs.}$$



www.nursunity.ml

2017